illumina®

Data Analysis: GWAS Processing

An overview of data processing using the BaseSpace® Correlation Engine for GWAS.

Introduction

The Illumina BaseSpace Correlation Engine contains over 25,000 biosets obtained by mining the vast amounts of publicly available genomic data from sources such as the Gene Expression Omnibus, ArrayExpress, and other repositories. These data go through a systematic screening, curation, and analysis process (Figure 1). This technical note describes the use of the BaseSpace Engine for analysis of data derived from genome-wide association studies (GWAS).

Biosets for the BaseSpace Engine are generated by processing the following types of source data from GWAS:

1. *Precomputed summary statistics using standard single-marker statistical analysis methods for GWAS.* These initial-phase biosets were generated from data analyzed from external sources (eg, from the database of Genotypes and Phenotypes [dbGaP][1]) to ramp up content for the sequence-centric application. However, steps were taken to make sure that they meet Illumina quality control criteria and significance thresholds. In the future, these studies can be reprocessed from raw data according to the data analysis protocol for GWAS described in this technical note. Criteria for inclusion of studies where only precomputed results are available include:
   a. Data collection and methods described in peer-reviewed journals
   b. Reasonable investigation into population stratification
   c. Quality control (QC) filtering for individual mutations and single-nucleotide variants (SNVs), or summary statistics (minor allele frequencies [MAF], Hardy-Weinberg Equilibrium [HWE] p-value, call rates) provided to apply standard QC filtering
   d. Valid association-testing methodology
5. *Literature-derived and curated SNV biosets from GWAS publications.* A list of GWAS studies was compiled from the National Institutes of Health (NIH) Catalog of Published Genome-Wide Association Studies,[2] excluding studies where data are already available via dbGaP.
6. *Raw data or genotype-level data*. This refers to data in the form of genotype calls or data in the form of raw files (eg, CELs) and when genotype calls are not made available. The following sections describe the analysis procedure used to create SNV biosets from the various sources of data. Standard statistical tests and analysis procedures[3-5] are used throughout the analysis protocol.

Raw Data Processing

Genotype calls are obtained from raw data using standard methods recommended by the platform manufacturer. For example, Chiamo[3,4] or BRLMM[5] can be used to obtain genotype calls from the 500k Affymetrix chip, and GenCall[6] can be used for raw data from Illumina chips. When raw data are already available in the form of genotype calls, they are used directly.

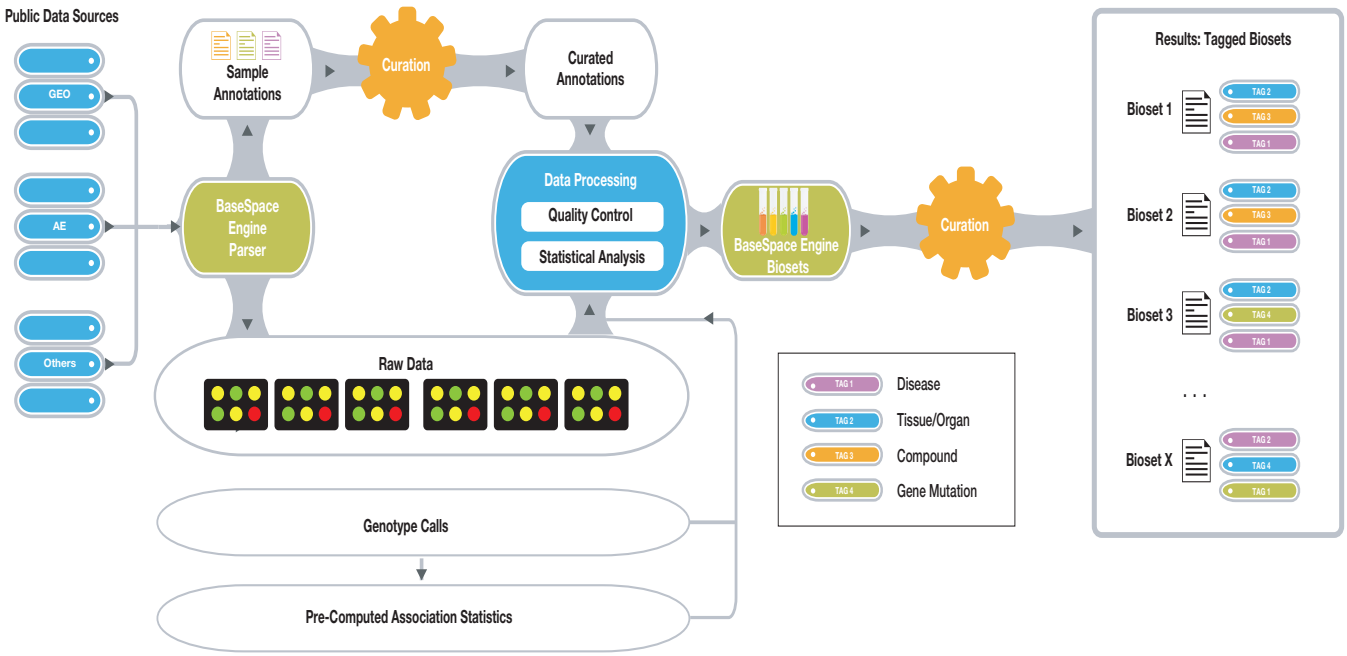

Figure 1: Public Data Analysis—The workflow for turning public data sets into processed gene biosets includes raw data collection, sample annotation curation, data quality control, automated analysis, and manual tagging of resulting biosets with disease, tissue, and compound ontology terms (tags).

Data Annotation

Study Annotation

Study annotation comprises:

- Study design
- Phenotype
- Sample inclusion criteria (eg, ancestry, clinical features)
- Study description

Sample Annotation

When raw or genotype level data are available, the parser extracts the following information for each sample from study annotation files:

- Family ID
- Sample ID
- Paternal ID (for family-based studies)
- Maternal ID (for family-based studies)
- Sex
- Phenotype (discrete or quantitative)
- Group/cluster (eg, geographical region) to assess possible effects of population stratification

Platform Annotation

When raw or genotype-level data are available, the parser extracts the following platform information:

- SNV ID (rs# or identifier)
- Alleles
- Chromosome
- Position (bp)
- Genetic distance (Morgans)

Exclusion Lists

When provided by the data source, exclusion lists for individual samples and SNVs are collated and parsed. For example, samples can be excluded for various reasons, such as:

- Discordance in genotyping when multiple platforms are used
- Discordance in clinical features (diagnosis, gender)
- Ancestry
- Duplications
- Related samples for population-based GWAS

Data Exploration and Generation of Summary Statistics

Genotype calls are analyzed and after basic preprocessing. Various summary statistics are generated before QC filtering as follows:

- Setting invalid genotypes to "missing"—eg, female Y genotype, heterozygous haploid chromosome
- Minor allele frequency (haploid chromosomes counted only one time)
- Missing genotype rate
- Missing rates by case or control status
- HWE failures
- Mendel errors (family-based data only)

QC Filtering

When whole-genome association statistics are available (precomputed or computed by the BaseSpace Engine), the following filtering criteria are applied to individual samples or markers.

Individual data are removed from further analysis if:

- Missing genotype data are > 10%
- Mendel errors are > 5% (family-based data only)
- Individuals are in an exclusion list
- Gender discrepancy exists between chromosome X data and reported sex

SNVs are discarded from further analysis if:

- MAF in both Cases and Controls *or* overall MAF is < 1%
- HWE p-value in Controls *or* overall HWE is $< 1 \times 10^{-6}$
- Average Call Rate in Cases and Controls *or* overall Call Rate < 95%
- Mendel errors are > 10% (family-based data only)
- SNVs are in an exclusion list

The filtering criteria applied to data sets curated from GWAS publications depend on the experiment type and are described in *Constructing a Bioset*.

Stratification Analysis

To investigate the possibly confounding effects of population stratification, the BaseSpace Engine employs the methods offered by PLINK[7] for complete-linkage agglomerative clustering, based on pairwise genome-wide identify-by-state (IBS) distance.

Multi-Dimensional Scaling

The BaseSpace Correlation Engine performs standard multidimensional scaling analysis on an N × N (N = total number of samples) matrix of IBS pairwise distances. Plotting the various dimensions against each other can be useful for identifying any clustering of samples. A typical visualization exercise is plotting the first dimension vs. second dimension and color-coding the individuals according to the cluster information (eg, ancestry and geographical location).

Genomic Control

An estimate of the genomic inflation factor (based on median chi-squared) is obtained using the Genomic Control method[8]. Adjusted test statistics are computed to correct for the genomic inflation factor.

Association Testing

Population-Based Association Testing

Case-control analysis

For all markers in the data set, multiple association tests are performed:

- Allelic association test
- Cochran-Armitage trend test
- Dominant gene action (1 degree of freedom [df]) test
- Recessive gene action (1 df) test
- Genotypic (2 df) test

Because there is generally no accepted answer to the question of which single-SNV test to use, the BaseSpace Engine adopts an intermediate choice. It determines the *maximum test statistic* from the tests above to cover an investigation into various effects: additive, dominant, and recessive.[9] However, when stratification information (eg, geographical location) is available, the Cochran-Mantel-Haenszel test for 2×2×K stratified tables is used instead. Association analysis statistics are visualized using Manhattan plots, histograms showing the p-value distribution, and Q-Q plots (Figures 2–4).

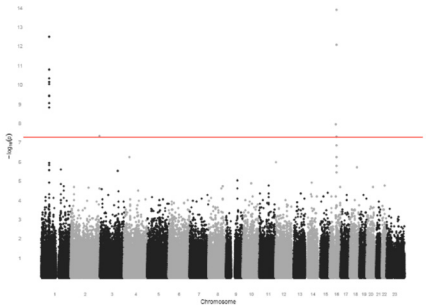

Figure 2: Association Analysis: Manhattan Plot—Manhattan plot of -$\log_{10}$ p-values of SNVs from association analyses of case-control samples in a GWAS study.

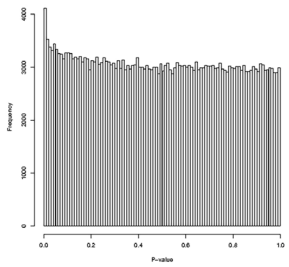

Figure 3: Association Analysis: Histogram—Histogram showing p-value distribution of SNVs from association analyses of case-control samples in a GWAS study.

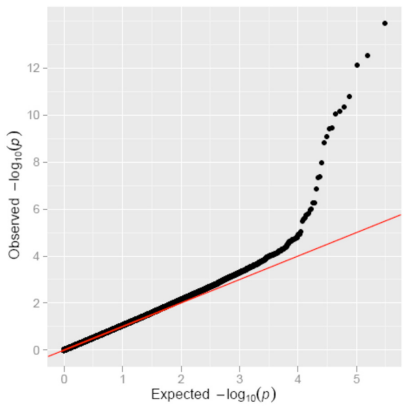

Figure 4: Association Analysis: Q-Q Plot—Log Quantile-Quantile (Q-Q) plot of observed vs expected -$\log_{10}$ p-values of SNVs from association analyses of case-control samples in a GWAS study.

Quantitative analysis

Quantitative traits are tested for association using the Wald test.

Family-Based Association Testing

Family-based association testing for disease traits is conducted using the standard transmission disequilibrium test (TDT) association test.

Adjusted Test Statistics

To account for the confounding effects of population stratification in family-based association studies, p-values are corrected using Genomic Control.[8]

To adjust for multiple testing, a false-discovery rate (FDR) analysis is used.[10,11] SNVs not meeting a minimum FDR (default q value = 0.2) are discarded from the bioset.

Constructing a Bioset

From Pre-Computed Analyses and Analyses of Genotype-Level or Raw Data

GWAS biosets contain significant SNVs that:

- Pass the QC Filtering criteria described above
- Satisfy a minimum FDR (default q value = 0.2).[10,11]

The following statistics associated with each SNV can be included in the bioset:

- Genotype call rates for cases and controls
- Genotype counts for cases and controls
- MAF for cases and controls
- HWE p-value for cases and controls, or combined HWE p-value
- Association test statistic
- Association test p-value (used for ranking SNVs within a bioset)
- Odds ratio with confidence intervals (default 95%)
- Risk allele
- Strand

From GWAS Publications

GWAS studies routinely consist of discovery and replication phases.

1. *Discovery Phase*: Genome-wide scans of thousands and hundreds of thousands of SNVs are employed to identify SNVs associated with the phenotype being studied. SNV biosets created from the discovery phase consist of top-ranked SNVs in the association study, or SNVs with p-value of 1×10^{-5} or less.
2. *Replication Phase*: Individual genotyping technologies are employed routinely to confirm the association, mostly in similar population cohorts and sometimes in different population cohorts. SNV biosets created from the replication phase consist of top-ranked SNVs in the association study or SNVs with p-value less than a cutoff determined by taking into account the platform size, if provided, or the number of SNVs reported as follows:
   a. If platform size is not available or number of SNVs reported < 500, set correction factor = number of SNVs reported.
   b. If platform size is available, set correction factor = platform size.
   c. p-value cutoff = max ((1×10^{-5}, 0.05) ÷ correction factor × 100), with the constraint that p-value cutoff is ≤ 0.05

3. *Combined Analysis*: Many articles report combined analysis of discovery and replication cohorts. In this case, a separate bioset for the combined analysis is created. Top-ranked SNVs satisfy a P value cutoff as follows:
   a. Set correction factor = number of SNVs reported.
   b. p-value cutoff = max ((1×10^{-5}, 0.05) ÷ correction factor × 100), with the constraint that p-value cutoff is ≤ 0.05.

An FDR-like analysis is also performed where the expected (by chance) and observed SNVs at various p-value thresholds (10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.05) are computed and visualized (Figure 3). The p-value cutoff is revised to the highest threshold that satisfies an expected/observed ratio ≤ 0.5, if it exceeds the p-value cutoff computed using steps 1–3.

BaseSpace Correlation Engine Upload

An upload package into the BaseSpace Engine consists of:

- A study description document, containing information about the study design, authors, data collection, platform, and phenotypes under consideration.
- Biosets containing significant SNVs with associated statistics. The bioset summary includes information about the composition, ancestry, number of samples in each cohort, platform used, and any testing that was performed to qualify participants (eg, affected vs. unaffected). It also contains analysis relevant to the generation of association p-values, population stratification, imputation, and metaanalysis. If the investigators have not applied imputation, then the NextBio platform imputes estimates of association statistics of untyped markers based on the reference population and platform used for the study.
- Bioset tags.

Imputation

When an SNV bioset is uploaded, the BaseSpace Engine can also estimate ranks or p-values for highly correlated SNVs that do not exist on the platform. Thus, the results of the data set being imported can be extended reliably to other searchable SNVs in the system. For SNV sets where imputation has not already been performed while processing, this is done as follows:

- SNVs outside the study platform that are in high linkage disequilibrium (LD) with SNVs in a bioset are assigned a rank (or P value) based on the ranks (or P values) of the highly correlated SNVs in the platform. Population-specific LD information (r^2) is obtained from HapMap 3.[12] Where applicable, the platform uses the pairwise correlation values pertaining to the reference population closest to the population under consideration in the study.
- If LD $r^2 \geq 0.8$, the rank of the imputed SNV is tagged as being estimated with high confidence.
- If LD $r^2 \geq 0.5$ but < 0.8, the rank of the imputed SNV is tagged as being estimated with low confidence.
- SNVs are mapped to genes where possible, and a gene-set enrichment analysis is performed to correlate the study to all the other BaseSpace Engine content.

Tagging Biosets with Ontology Terms

The final step in preparing data for use in the BaseSpace Correlation Engine is the tagging of biosets with standard ontology terms reflecting disease, tissue, treatment, experimental design, and sample source (Figure 1). This step is important and provides a semantic backbone to all the analyzed data. This enables the BaseSpace Correlation Engine to perform additional high-level correlations and also makes it easier for users to find relevant data sets. Comprehensive ontologies cover anatomy (FMA/MeSH),[13] diseases (SNOMED CT), and compounds (over 8 million compound clusters and associated structures).

Ranking of SNVs

SNVs within a bioset are ranked by p-value. The absolute rank value is computed from a transformation function that enables more accurate comparison of SNV biosets to biosets from other data types. Details of the ranking metrics are provided in the technical note *Ranking of Genes, SNVs, and Sequence Regions*.

References

1. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007;39(10):1181–1186.
2. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA*. 2009;106(23):9362–9367.
3. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007;447(7145):661–678.
4. Marchini J, Howie B, Myers S, McVean G, Donnelly P A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*. 2007;39(7):906–913.
5. media.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf
6. support.illumina.com/downloads/gencall_software.ilmn
7. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–575.
8. Devlin B and Roeder K. Genomic control for association studies. *Biometrics*. 1999;55(4):997–1004.
9. Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet*. 2006;7(10):781–791.
10. Storey JD A direct approach to false discovery rates. *J R Stat Soc*. 200264(3):479–498.
11. Storey JD and Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*. 2003;100(16):9440–9445.
12. International HapMap Consortium. The International HapMap Project. *Nature*. 2003;426(6968):789–796.
13. sig.biostr.washington.edu/projects/fm/

Illumina • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

illumina®